

Criteria for locally dense subgraphs

Gergely Tibély

Institute of Physics and HAS-BME Cond. Mat. Group,
Budapest University of Technology and Economics,
Budapest, Budafoki str. 8., H-1111

Abstract

Community detection is one of the most investigated problems in the field of complex networks. Although several methods were proposed, there is still no precise definition of communities. As a step towards a definition, I highlight two necessary properties of communities, separation and internal cohesion, the latter being a new concept. I propose a local method of community detection based on two-dimensional local optimization, which I tested on common benchmarks and on the word association database.

1 Introduction

In the last decade interdisciplinary research on complex networks resulted in spectacular development [1]-[6]. It has become clear that networks constructed from diverse complex systems show remarkably similar features. Several aspects were investigated, like clustering [7], the degree distribution [8], diameter [9], [10], spreading processes [11], diffusion [12], synchronization [13], critical phenomena [14] and game theoretical models on complex networks [15].

One of the most actively researched questions about complex networks is the one of community detection [16]. Community detection aims at finding dense groups in graphs, like circles of friends in social networks, web pages about the same topic, or substances appearing in the same pathway in metabolic reaction networks. Perhaps the strongest motivation behind the research is that dense groups in the topology are expected to correspond to functions performed by the network, such that one can infer from pure topology to function. While the concept of communities seems intuitively plausible, attempts for an algorithmically useful definition have not been successful yet. The global characterization by modularity [17] or by random walks [19, 18], the local “weak” and “strong” definitions [20], the clique percolation approach [21], or the multiresolution methods [22, 23, 24] have all increased our understanding of this complex problem but the proliferation of methods of community detection just indicates the difficulty of this issue [16].

Unfortunately, any precise definition of communities is still lacking, giving rise to innumerable methods using different definitions. Lack of a definition also makes problematic the testing of methods; although there is progress in this issue [25], [26]. Difficulty of the problem is increased by more subtle factors: very

often communities occur on a broad scale, they can be ordered in a hierarchical manner, and they may overlap, which make their identification even harder.

After being the subject of active research for several years, it is getting clear that the following stages appear during community detection:

- 1 defining the term “community”;
- 2 finding the objects corresponding to the definition;
- 3 determining the significance of the found communities.

Although from the theoretical perspective stage 1 is clearly a key issue, it is far from being settled. Several different propositions exist, which are evaluated mostly according to their results on a few benchmarks. This is the stage to be improved in the first place in this paper. Stage 2 is a technical issue, often consisting of some combinatorial optimization method. Its choice is usually a result of a trade-off between speed and quality. Stage 3 should give information about how surprising is the existence of a found community in the actual graph, given some characteristics of the graph like edge density or degree distribution. Although this issue also got some attention [27]-[33], it just began to get widespread application [34].

The rest of the paper will focus on the question of definition, so a few remarks about stage 3 are made here. Most community detection methods give no information about the significance of their output, thus forcing the investigator to assume that all results are (equally) significant. This way, the community detection stages 2 and 3 are combined into a single decision whether a particular subgraph is a good enough community or not – effectively pruning the significance test in practice. The other end of the spectrum, represented by [34], builds the definition of communities on statistical significance, which is clearly an improvement. However, it should be noted that the fitness and statistical significance of a subgraph as a community are not synonyms. Statistical significance tells us how surprising a subgraph is, while fitness talks about how close is it to the ideal community. Therefore, the two quantities are complementary and both belong to the description of a community.

2 Local criteria for communities

A fundamental problem of community detection is to define the term “community”. There are different approaches to this question. One is the algorithmic approach, giving a computational procedure for finding clusters. This naturally incorporates a mathematically precise definition, although different algorithms usually result in diverse definitions, and there is no theoretical framework currently to help their differentiation. Another possibility is to present a general concept, on which a precise definition can be based. In this paper, the latter approach is taken, although an algorithmic realization is also presented.

No definition of communities which is both precise and generally accepted has appeared yet. Currently the description of communities exhausts in the phrase “nodes having more edges among themselves than to the rest of the graph” (or equivalent forms). It can be translated roughly to “statistically significant locally dense subgraphs”. Statistical significance is a quite precise expression, the main problem is with the term “locally dense”. For an intuitive

picture, it is quite good, but much less than directly transformable to algorithms. Although there is an implicit agreement on that clearly counterintuitive results are not permitted, even a formal list of required properties is missing. However, there are some properties which fit human intuition about locally dense subgraphs¹²:

Separation: a good community is well-separated from the rest of the graph;

Cohesion: a good community is homogeneously well connected inside, i.e. it is hard to separate into two communities.³

The separation criterion is quite clear, although there is an important remark: separation should be defined locally, involving only the community under investigation and its immediate neighborhood. Global methods, in which distant regions of the graph can modify a community in order to improve a global fitness value, can produce results violating the human perception about clusters. A famous example is the resolution limit of modularity [35, 36].

Although separation is a very intuitive criterion, and famous methods rely on it (see the Appendix), it is not enough in itself. Figs. 1a-1b illustrate that the distribution of links inside the separated region (the “shape“ of the subgraph) also matters heavily. Application of current community detection methods to real-world networks confirms that this is a real problem, e.g. tree-like communities can occur, even when the whole network is not tree-like [37], [38].

Both separation and cohesion are required properties of communities. If one neglects cohesion, the result may contain clusters like the one on Fig. 1a. On the other hand, if separation is not taken into account, one may end up chopping a separated subgraph until very cohesive pieces (cliques in the extreme) are obtained, like the triangles on Fig. 1b.

Given the subgraphs on Figs 1a and 1b as proposed communities, most community detection methods’ fitness values, to be reviewed in the next section, can not tell the difference between them. This is due to that most methods simply count the internal and/or external edges, which do not tell about the distribution of those edges. The reason why several methods do not fail to assign proper clusters for Fig 1a is that they look for optimal clusters, consequently they compare configurations like Fig 1a in one cluster and in two clusters, and splitting the two cliques into two clusters may improve the partition. But the situation is even worse. In the next Section, we will see that a number of fitness functions are more optimal for a counterintuitive clustering than for the intuitive one (e.g. joining the two cliques on Fig 1a, like modularity for a large enough graph). It should be noted that in such a case, the proper communities might be recovered if the heuristic gets stuck in the proper local optimum, even when that is not the global optimum.

¹For brevity, the words “community”, “group” and “cluster” will be used from this point as synonyms for “locally dense subgraph”, omitting the statistical significance from the meaning.

²It should be noted that the meaning of the term “community” can depend on the context; consequently a single definition may not be enough. Here the aim is to describe a particularly intuitive one.

³The term “cohesion” also appeared in [22], although there it denotes a quantity with an unrelated concept.

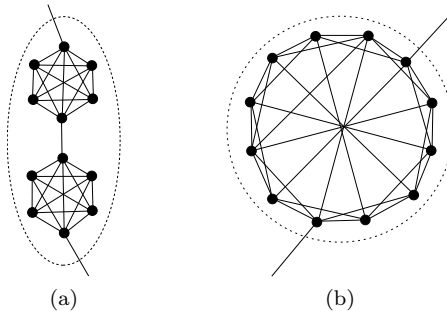


Figure 1: Illustration of the importance of subgraph shape. The two subgraphs have the same number of nodes and the same degrees, i.e. they differ only in the distribution of links. The figure on the left is much less cohesive than the figure on the right, although just a reorganization was applied to the links.

3 Overview of the existing methods

Here, existing community detection methods will be reviewed from the point of view of the previous section, i.e. do they conform the criterions of separation and cohesion. As mini-benchmarks, the examples on Fig. 1 or their simple variations will be used (see the Appendix for details on individual methods). The desired output for 1a is two communities consisting of the two cliques, while 1b should be kept in one piece. In both cases, no nodes from the rest of the graph should be included. For methods optimizing a fitness function, the globally optimal solution will be considered, for other methods, the possible solutions. These solutions will be compared to the desired ones, independently for Fig. 1a and 1b. If a method separates the two cliques of Fig. 1a, then it gets a “+”, if it puts all nodes of Fig. 1b into one cluster, then it gets another “+”. If there are multiple equally valid solutions (like for label propagation), all solutions are required to conform the preferred result.

For methods optimizing a function, the heuristic realizing the optimization may deviate from the global optimum, presenting worse or even better results (in terms of conformity to separation and cohesion). This will not be investigated, here the focus is on the definition of the communities (following from the choice of the fitness function), not on the practical aspects. Results for methods which can produce a single partition or cover are displayed in Table 1. The large number of published methods makes assembling a complete list nearly impossible. Instead, the emphasis is put on the diversity of the reviewed approaches.

There is a bunch of multiresolution methods, which possess a parameter allowing to tune the cluster sizes from 1 (isolated nodes) to $\mathcal{O}(N)$: the multiresolution modularity of Reichardt and Bornholdt (RB) [22], of Arenas, Fernández and Gómez (AFG) [23], the local fitness method of Lancichinetti, Fortunato and Kertész (LFK) [39], the Potts model of Ronhovde and Nussinov (RN) [43], the Markov autocovariance stability of Delvenne, Yaliraki and Barahona (MAS) [19], the hierarchical likelihood method of Clauset, Moore and Newman (CNM) [56], and the Markov Cluster Algorithm of van Dongen (MCL) [57]. Naturally, these methods are expected to find the proper community assignments both to

method	cohesion test (like Fig. 1a)	separation test (like Fig. 1b)
Lancichinetti et al. [39]	-	+
Labelpropagation [40]	-	-
Infomap [18]	-	+
Clique Percolation [21]	-	-
Estrada & Hatano [41]	-	-
Modularity optimization [17]	-	+
Donetti & Muñoz [42]	-	+
Ronhovde & Nussinov [43]	-	-
Nepusz et al. [44]	+	-
Hofman & Wiggins [45]	-	-
Hastings [46]	-	-
Newman & Leicht [47]	+	-
Wang & Lai [48]	+	-
Bickel & Chen [49]	+	-
Karrer & Newman [50]	+	-
Infomod [51]	-	+
Radicchi et al. [20]	+	-
Chauhan et al. [52]	+	-
Evans & Lambiotte [53]	-	+
Ahn et al. [54]	-	-
ModuLand [55]	-	-

Table 1: Cohesion & separation criterion test results. Tests were done on Fig. 1a and 1b or similar graphs (which are described in the Appendix). + and - are assigned according to whether the fitness function of a method is more optimal for the preferred solution or not. For methods which do not optimize a fitness function, simply the possible solution(s) was (were) analyzed. See the Appendix for details on specific methods.

Fig. 1a and 1b at some parameter values. However, there is no guarantee that these values are also the proper ones for the rest of the graph. Consequently, it is not clear how a resolution parameter should be set: the natural idea is to find the longest interval of the resolution parameter value in which the community structure does not change, but when the optimal parameter value is different for different regions in the graph, the longest stable interval not necessarily reflects the optimal communities.

Furthermore, the fitness values do not help us to tell good clusters from bad ones, like Fig. 1a from Fig. 1b. For most multiresolution methods (RB, AFG, LFK, RN), it is very easy to see that the fitnesses of two clusters are the same given that all nodes has the same in- and outdegrees, independently of the shape of the clusters. Note that it is also true for most single resolution methods. For MAS it is not trivial. Therefore, empirical tests were conducted to check it. According to them, Fig. 1a was found empirically to be at least as good as

1b⁴. Finally, regarding MCL and CNM, they have no fitness function⁵, the only accessible quantity about the community structure is the parameter interval in which it is stable.

Finally, there are hierarchical methods, which look for series of smaller and smaller (or larger and larger) clusters hierarchically embedded into the previous ones. Similarly to multiresolution methods, they are expected to contain good clusters in the outputted hierarchy. However, when looking at a graph having a simple one-level community structure, the question how to select the proper levels of the outputted hierarchy arises. The easiest way is to use the lowest level communities. Unfortunately, it is not a reliable procedure, as the lowest-level clusters may be just parts of the communities of the optimal partition or cover (see the Appendix for details). A second idea can be to assign significance scores to the communities on different levels, in the spirit of [29]. Although this approach might reliably qualify the found communities, a new version of statistical significance taking into account the internal cohesion is required. Furthermore, one should be very careful not to impose unnecessary constraints, like prohibiting overlaps, when constructing a hierarchical method.

A further question is whether a method provides information about the shape of the found communities or not. Recent analysis of real-world networks highlights the relevance of this issue [37], [38]. Several methods are based on simply counting the internal and/or external edges, or degrees at most: LFK, Labelpropagation, Infomap, modularity optimization (and equivalents), Hofman & Wiggins, Hastings, Ronhovde & Nussinov, Newman & Leicht, Wang & Lai, Bickel & Chen, Karrer & Newman, Infomod, Ahn et al., OSLOM. Consequently, they do not see any difference in the distribution of the links, e.g. Fig 1a and 1b get the same fitness values. Only Clique Percolation and Radicchi et al.'s⁶ method have some very limited requirement about cohesion built in the definition of communities.

The conclusion is that none of the reviewed methods is able to successfully apply both the separation and the cohesion criterions. They susceptible either to glue together well-separated subgraphs or to overpartition a cohesive subgraph. Future network designs should consider cohesion as well as separation.

4 Community detection in a two dimensional parameter space

In this Section, a new method for community detection is introduced. Its main goal is to present a method which takes into account both criterions defined in Sec. 2. First, the LFK method will be reviewed, which will serve as a starting point for the new method. Then, a composite fitness will be constructed which takes into account the separation and cohesion criterions. Finally, a heuristic optimization procedure for the composite fitness will be described, which finds

⁴In this case, only 1 link to the rest of the graph was used. Rest of the graph, represented by a single node having self-loops, was assigned 118 edges inside, resulting in a total of $L = 150$ edges. Stability values were calculated from 0.01 to 100, the step size being 0.01 below 1.0 and 1 above.

⁵CNM does have a fitness function, but it corresponds to a full hierarchical dendrogram, not to any partitions obtained by cutting the dendrogram at some point

⁶If the stopping criterion of their heuristic is considered as part of the definition.

locally dense subgraphs on all scales, and also able to recover hierarchical structures.

4.1 The LFK method

The LFK method [39] optimizes the local fitness function

$$f^C = \frac{K_{in}^C}{(K_{in}^C + K_{out}^C)^\alpha} \quad (1)$$

where C denotes a subgraph, K_{in}^C and K_{out}^C are the total number of inside and outside degrees in C , respectively, and α is a tunable exponent for setting the size scale of the communities to be found. Running the method with large α values result in small clusters, small values in large clusters. The recommended range for α is 0.5-2.

The practical implementation of the optimization works as follows. The communities are found one-by-one, independently of each other. First, a seed node is selected from which the new community will be grown. Then, the node which can best improve the fitness of the cluster is added. This addition is repeated until the fitness reaches a local optimum. After each addition, removal of nodes takes place, if the fitness can be enhanced that way. When the fitness cannot be further increased, the actual subgraph is declared a community. The growth process is repeated for all nodes as seeds, or alternatively, until the found communities cover all nodes in the graph.

Although the resolution parameter α can be tuned continuously, [39] suggested that the relevant community structures should be identified by robustness to changes in α , i.e. which have the longest interval for α values without change. Changes in the community structure were detected by monitoring the mean fitness of the communities, evaluated at a reference value $\alpha = 1$.

4.2 Implementing the criterions

For the separation criterion, the following function will be applied

$$f_S^C = \frac{K_{in}^C}{K_{in}^C + K_{out}^C} \quad (2)$$

where C is a subgraph, K_{in} and K_{out} are the sums of in-community and out-community degrees, respectively. This is the fitness of LFK [39], with the multiresolution parameter being set to one. For detecting hierarchical structures, a different solution will be described. Eq. 2 clearly focuses on the external separation of the clusters, therefore it is suitable as an implementation of the first criterion of the communities.

For the internal cohesion criterion, a possible solution is to consider the second eigenvalue of the Laplacian matrix of the community. The Laplacian of a graph is the matrix $L = A - D$, where A is the adjacency (or weight) matrix, and $D = \text{diag}(k_i)$ is a diagonal matrix containing the degrees (strengths). Its largest eigenvalue is always 0 (corresponding to the trivial eigenvector $(1, 1, \dots, 1)$). The multiplicity of the largest eigenvalue equals to the number of connected components in the graph. This gives the hint that if two distinct graphs are got connected by a single (weak) link, the Laplacian gets only a slight perturbation

(compared to the case of two connected components), which splits the double degeneracy of the first eigenvalue, such that a new eigenvalue close to zero appears⁷. In fact, it is known that the second eigenvalue of the Laplacian measures “how difficult is to split the graph into two large pieces” [58].

For some important special cases the second eigenvalue can be calculated:

- for full graphs of n nodes (clique), $\lambda_2 = -n$
- for a star-graph, $\lambda_2 = -1$, independently of n
- for a linear chain, $\lambda_2 = -2 + 2 \cos(\pi/n) \rightarrow 0$ as $n \rightarrow \infty$
- for two n -sized cliques attached by a single link (having weight ϵ) (like on Fig. 1a), $\lambda_2 \approx -\frac{2\epsilon}{n+2\epsilon}$, which also goes to 0 as $n \rightarrow \infty$.
- for a disconnected graph, $\lambda_2 = 0$. This may seem trivial, but most methods give a finite score for disconnected communities; it is not without precedent that such objects can be produced in reality [38]. Although this problem can be avoided by a properly designed heuristic of a method, disconnected communities should be punished by definition.

Calculation for the two cliques is in the Appendix, other results can be found in [59]. These cases confirm that the second eigenvalue is useful for quantifying the cohesion criterion of the definition of communities. For an illustration, on Fig. 2 a few example graphs with their second Laplacian eigenvalues are shown.

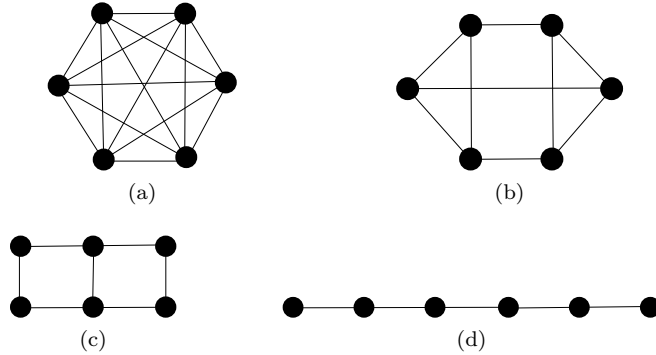


Figure 2: Graphs with different second Laplacian eigenvalues. $\lambda_2^{(a)} = 6$, $\lambda_2^{(b)} = 2$, $\lambda_2^{(c)} = 1$, $\lambda_2^{(d)} = 0.268$. The maximal value of λ_2 is 6 in all cases.

The separation fitness term f_S ranges from zero to one. In order to compose it together with the cohesion fitness, the latter should also be in the interval $[0, 1]$. Therefore, λ_2 needs some transformations before application as fitness. As can be seen from the above examples, for the worst cases $|\lambda_2|$ is of the order of $1/n$, therefore the lowest point of the $|\lambda_2|$ -scale will be set to $1/n$. The highest point is trivially given by n . It is reasonable to assume that most subgraphs have $|\lambda_2| = o(|C|)$. Furthermore, several subgraphs can have worse internal cohesion

⁷The diffusion matrix was also considered, but it prefers star-like graphs too much.

than the star graph, thus having $|\lambda_2| \in [0, 1]$. To take into account these effect, $\log |\lambda_2|$ will be more useful than λ_2 . So, in order to obtain a quantity between 0 and 1, the minimum will be subtracted and divided by the maximum,

$$f_C^C = \frac{\log |\lambda_2| - \log 1/|C|}{\log |C| - \log 1/|C|} = \frac{1}{2} + \frac{1}{2} \frac{\log |\lambda_2|}{\log |C|}, \quad \text{if } |C| > 1$$

$$= 0 \quad \text{if } |C| = 1 \quad (3)$$

where $|C|$ is the number of nodes in the community. The above measure happens to be 0.5 if λ_2 is 1, e.g. for the star-graph. I wish to emphasize that eq. 3 is only one possible proposition for taking into account the internal cohesion, although a promising one – better measures may exist. The same is true for the choice of f_S .

The cohesion fitness f_C^C opens the way for constructing tests assessing the performance of community detection methods regarding the cohesion of the found communities. One may generate a graph with built-in communities which separation is controlled, like in the LFR benchmark [25], then randomly select pairs of clusters and increase the interconnection between the two members of each pairs to some predefined value, finally calculating f_C^C of the pairs. Running the detection method and measuring the ratio of pairs not split as a function of f_C^C may indicate how strongly focuses the method on cohesion.

The next question is how to combine $f_{separation}^C$ and $f_{cohesion}^C$. Thinking in a two dimensional space of f_S^C and f_C^C , a natural approach is to get as far from the point (0,0) as possible. This implies

$$f^C = \sqrt{(f_S^C)^2 + (f_C^C)^2} \quad (4)$$

so the fitness is the euclidean distance from (0,0). Again, this is just one possibility, better combinations may exist. E.g. the relative weight of f_S and f_C may be adjusted in a more well-grounded way. However, eq. 4 is able to pass the test raised by Fig. 1: for Fig. 1a, $\lambda_2^{2\text{cliques}} = 0.258$, $f_C^{2\text{cliques}} = 0.228$, $f^{2\text{cliques}} = 0.995$ while for a single clique $\lambda_2^{1\text{clique}} = 6$, $f_C^{1\text{clique}} = 1$, $f^{1\text{clique}} = 1.371$. For Fig. 1b, $\lambda_2^{12\text{nodes}} = 3.268$, $f_C^{12\text{nodes}} = 0.738$, $f^{12\text{nodes}} = 1.218$, and for the best subgraph, a triangle, $\lambda_2^{\text{triangle}} = 3$, $f_C^{\text{triangle}} = 1$, $f^{\text{triangle}} = 1.077$.

Beyond enabling one to decide whether a given subgraph is a community or not (by requiring local optimality), the above definition makes it possible to assess how good community it is. This is also possible with another definitions, e.g. by using the modularity function, but here, communities are placed on a 2-dimensional space instead of 1 dimension. This gives rise to an interesting possibility for characterizing the communities, like “very cohesive but densely connected outwards” or “well-separated but poorly interconnected”. Considering Fig. 1a, one may think that the latter is not really a community. But for large subgraphs, it may make sense to consider a well-separated subgraph as a community, as common sense says that large communities should be looser than small ones.

4.3 Community detection in reality

In this section, the details of practical implementation of the new method are discussed. Most importantly, in order to actually find the communities, a heuristic

carrying out the optimization of eq. 4 is needed. Furthermore, there is a second problem of detecting communities hierarchically embedded into each other. These two questions will be answered by a common solution.

The heuristic is based on the one of the LFK method [39]. Among its details, the LFK heuristic contains a tunable parameter (denoted as α), which is claimed to be able to recover communities at different hierarchical levels. Lowering this parameter α results in increased community sizes. Hierarchical levels are supposed to be stable against the variation of α , so there should be long intervals for α for which the communities do not change. However, large graphs may lack long stable intervals, as some changes occur around any parameter value (data not shown). Therefore, a new method for investigating hierarchical structure is needed. I dropped the idea of using threshold values of α , corresponding to community structures at different scales, which should be simultaneously valid for all communities, and I will treat each community separately.

Similarly to [39], each community is grown from a seed node. It is important to note that each seed node can result in a series of (successively larger) communities. Growth consists of successively including the neighboring node which increases most the fitness defined by eq. 4. When there is no neighboring node which inclusion can improve the fitness, the stage of node removal begins. Here, the fitness of the cluster is tried to be improved by excluding nodes from it (with the exception of the seed node, which is not permitted to be excluded). It finishes when no further removal can improve the fitness. Then, growth begins again, if possible. The grow-shrink cycle is iterated, as long as the fitness can be improved. When no improvement is possible (there is a local optimum of the fitness), the actual list of nodes is registered as a valid community. After that, the algorithm tries to find a larger community, which contains the current one. This way, hierarchical structures can be revealed. In order to do it, first the growing cluster should escape from the basin of attraction of the current local optimum. Therefore, the cluster is forced to grow, by successively including the neighboring nodes which decrease the fitness the least. After some steps of forced growth, when increasing the fitness becomes again possible, the algorithm turns back to the normal grow-shrink procedure, until a new local optimum is found, signing a new community. The cluster keeps hopping from local optimum to another local optimum until it grows so large that it contains the whole graph. Then a new growth process starts from a new seed node. At the end of its growth process, it includes the whole graph again, unless it encounters a local optimum which has been already found, i.e. the corresponding community has already been registered. In this case, the growth process is stopped. Then, another growth process starts from a not-yet-used seed node. In contrast to [39], all nodes in the graph are used as seed nodes, in order not to miss good communities. When the growth process beginning from the last seed node finishes, the algorithm ends, and the registered communities are written to the output. There are a few additional tricks. First, if escaping from a local optimum seems to be hard, i.e. after changing from forced growth to the normal grow-shrink stage we still end up in the previous local optimum, the cluster is restored to the state where it had its maximal size (the beginning of one of the removal sessions), then 2 steps of forced growth is applied before the normal grow-shrink cycle begins. A second trick is that when judging the identity of two communities, they are considered identical if at least 80% of the larger community is

a subset of the smaller one⁸. In case of identity, the community which has the higher fitness is kept in the registry.

The algorithm, although based on the one of [39], differs in several points: from one seed, several communities can be reached instead of only the smallest one; node removal occurs when node addition is not possible instead of after each addition (this trick also speeds up the algorithm); seed node is not permitted to be removed; all nodes are used as seeds instead of the not-yet-covered nodes. An algorithm similar in spirit was described in [60]. The results in the next section are obtained using this method, unless stated otherwise explicitly. The software realizing the algorithm is available at <http://www.phy.bme.hu/~tibelyg/>.

5 Test results

Probably the most frequently used test is Zachary’s karate club friendship network[61]. Due to a dispute between two prominent persons (node 1 and 34), the club split into two during sociological observation, and the memberships in the new clubs are known. As the split occurred more or less along a border of two visible communities, new community detection algorithms are usually claimed to pass the test if they reproduce the split. However, the aim is the detection of *topological* modules, not functional ones, so the result of the sociological study is not a strict criterion for judging the output of any community detection method. E.g., node 10 has 1 – 1 links to each of the new clubs, so “misplacing” it (compared to the split) may not be considered as a fault. Or node 12, which attaches only to node 1, is hard to be considered as part of a “densely interconnected” cluster.

The algorithm finds 33 groups, containing several non-relevant ones, like pairs of nodes. Therefore, a filtering procedure is required. The statistical significance of the resulting communities [29, 34] is utilized for this purpose. The statistical significance can be sensitive for missing nodes [29], therefore each cluster is allowed to be completed with the neighboring node which optimizes the statistical significance. Then the clusters are ordered according to their statistical significance. The first 3 clusters provide a single-level community structure, corresponding to 3 known communities, with 2 overlapping and 1 homeless nodes (Fig. 3, left panel). Taking a look at the subsequent clusters provides information about the multi-scale structures in the graph. The next few clusters reveal cluster cores and hierarchical decomposition of the network (Fig. 3, right panel). The statistical significance score is quite capable of distinguishing meaningful structures; there is a gap between 0.42 and 0.81, so setting a threshold to 0.5 selects the multi-scale clusters which would be approved by a human investigator. There is only one exception, the almost-full-clique of nodes {1, 2, 3, 4, 8, 14} has significance 0.81, which is probably the consequence of neglecting the internal cohesion by the current form of statistical significance.

The currently most advanced class of benchmarks was introduced by [25]. In these so-called LFR benchmarks, the network size and edge density are freely adjustable, and more importantly, the node degrees and the community sizes are distributed according to power-law distributions, with tunable exponents.

⁸If the criterion were based on some percent of the smaller group, subset-superset pairs would be considered identical.

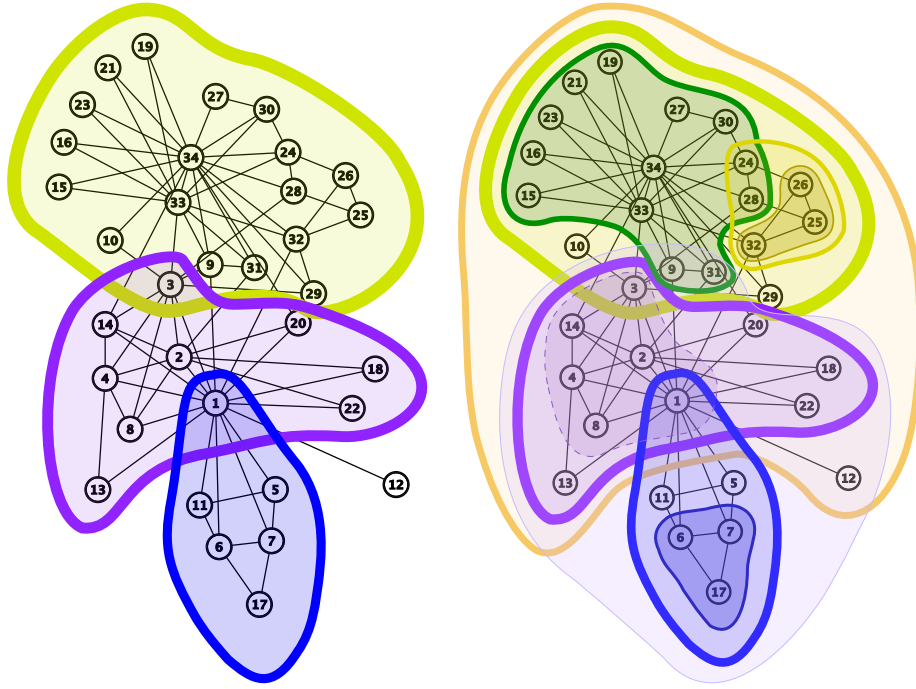


Figure 3: (Color online) The 3 (left) and 10 (right) best found communities of the Zachary karate club. On the right, thicknesses of lines indicate the ordering of the statistical significance values (running from 0.002 to 0.42, plus 0.81 for the dashed line-bordered community). Note that node 12 is contained only by large communities.

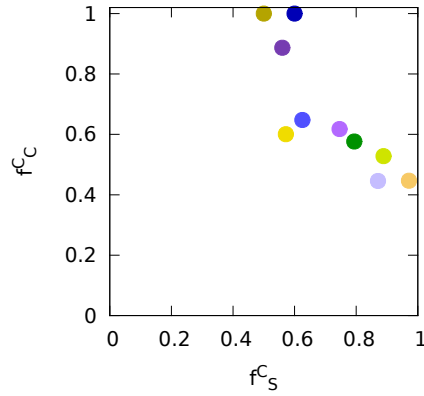


Figure 4: (Color online) Positions of the Zachary communities on the f_S - f_C plane. Small groups tend to cluster at North, and large groups at East.

Communities are defined through a prescribed ratio of inter-community links for each node (mixing ratio, μ), similarly to the preceding GN benchmark class [17]. Generalizations for weighted and directed networks, and for overlapping communities also exist [26].

A wide-scale comparison of different community detection methods using the LFR benchmark was done by [62]. For the ease of comparison, the parameter values of [62] are applied here: the networks consist of 1000 nodes, the average degree is 20, the maximal degree is 50, the exponent of the degree distribution is -2 and the exponent of the community size distribution is -1. There are two types of networks, for the S type the community sizes are between 10 and 50 (“small”) and for the B type they are between 20 and 100 (“big”). In [62], networks of 5000 nodes were also investigated. Due to the large computational time, they are omitted here⁹. Also for computational time considerations, the detecting algorithm stopped growing the communities over a predefined size, 120 for the S case and 220 for the B case. All measurement values are obtained from runs on 10 different networks.

Similarity of the built-in and the obtained community structures are quantified by a variant of the normalized mutual information (NMI), which is able to handle overlapping communities [39]. This is the similarity measure applied by [62]¹⁰.

Selecting the most relevant communities from the abundant output was done similarly to the previous case. The clusters were completed by 1 neighboring node, if that improved the statistical significance, and sorted with respect to the statistical significance scores. The clusters containing at least 1 uncovered node were accepted one by one until all nodes were covered.

To see the potential of the new method, and check the effect of the output-filtering, the communities corresponding best to the built-in original ones were also selected from the algorithm’s output. The results are plotted on Fig. 5 (a). The filtered results are similar to the ones of the lower performing algorithms in [62], while optimal selection provides much better scores, although still not as good as the best methods. The large difference between the optimal and the statistical significance-based results is quite surprising, especially in the light of the fact that statistical significance in itself is able to provide excellent results on the LFR benchmark [34].

The algorithm was also tested on networks with overlapping communities. In this case, clusters having significance score below 0.1 were accepted, similarly to [34]. Fig. 5 (b) shows that the effect of the imperfect output-filtering is again very large, an ideal selection scheme would allow very good results. This is not surprising, as other algorithms based on the one of [39] also give excellent results on overlapping communities [63].

Finally, the new method was applied to a word association graph built from the University of South Florida Free Association Norms [64]. Here, nodes are words and edges show that some people associated the corresponding two words.

⁹it does not mean that a single 5000-sized graph is too large, however, a few hundred of them are

¹⁰Both the LFR benchmark and the generalized normalized mutual information are freely available from the authors’ web-sites, <http://sites.google.com/site/santofortunato/inthepress2> and <http://sites.google.com/site/andrealancichinetti/software>

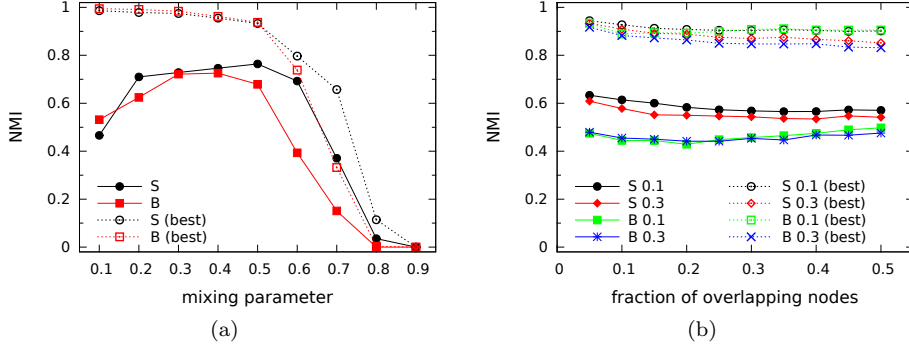


Figure 5: (Color online) Results on the LFR benchmark. Panel (a) corresponds to unweighted, undirected and non-overlapping tests, while panel (b) corresponds to overlapping tests. Overlapping tests were done at two different values of the mixing parameter, at $\mu = 0.1, 0.3$. For both panels: full symbols and lines correspond to the applied filtering, empty symbols with dotted lines correspond to perfect output filtering.

The network has 5018 nodes with mean degree $\langle k \rangle = 22.0$. It is a frequently used example of overlapping community structure [34], [21]. Although edge weights are accessible, the algorithm was applied to the unweighted version of the network. As an illustration, low-level communities around the word *bright* are plotted on Fig. 6. An interesting effect is the appearance of *overlapping edges*, due to the heavy overlap in the network.

In conclusion, although selecting the relevant communities from the output is not an already solved task, the algorithm gives good results on the Zachary karate club, and performs reasonably on the LFR benchmarks. It should be noted however, that due to the internal cohesion criterion, this algorithm’s output is not intended to perfectly match benchmarks like GN and LFR, which define communities solely on the basis of external separation. An additional observation is reported here: on GN benchmark graphs¹¹ with nodes having exactly the prescribed in- and out-degrees, at large mixing ratios communities deviating from the built-in ones but having better-than-designed mixing ratios were found. Note that the new method does not optimize just for external separation, so even better “spontaneous” communities may exist. This phenomenon, although not being a huge surprise, raises the question how to judge precisely a community detection method’s output at large mixing ratios, as the known community structure may not be trusted to 100%.

6 Discussion & Conclusions

An important aspect of all community detection methods is the *running time*. In the case of the new method described above, the time requirement is as follows. Starting a new community from each node contributes a factor of N to the CPU

¹¹results are omitted, as the presented LFR benchmark is a generalization of the GN.

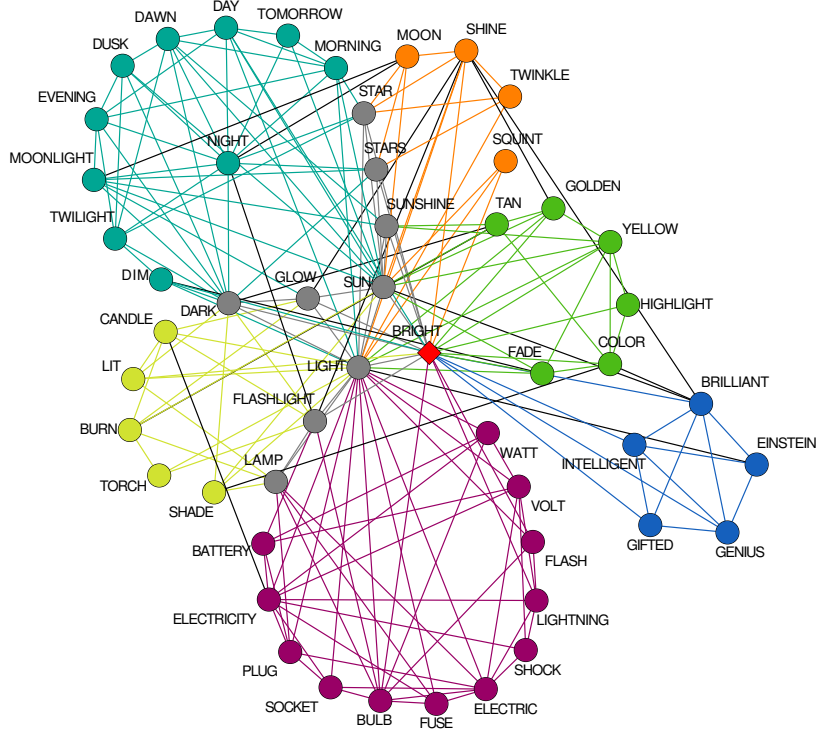


Figure 6: (Color online) Communities around *bright*, on the first hierarchical level. Color denotes communities. Gray shows overlapping nodes and edges. Black edges are between different communities.

time. Evaluating the eigenvalues of a community C plus one extra node takes $2/3(|C| + 1)^3$. Assuming that C has $\text{const} \cdot \langle k \rangle \cdot |C|$ neighboring nodes (i.e., on average, each node has a constant fraction of its neighbors outside C), running time can be estimated as

$$T \approx N \cdot \sum_{|C|=1}^{|C|_{\max}} \text{const} \cdot \langle k \rangle \cdot |C| \cdot (|C| + 1)^3 \approx N \cdot \text{const}' \cdot \langle k \rangle \cdot |C|_{\max}^5 \quad (5)$$

A naive estimate for $|C|_{\max}$ would be N . However, as more and more community growing processes finish, the newly started communities are expected to terminate in a previously discovered community earlier and earlier, on average. Of course, some communities will reach $|C| = N$. Therefore,

$$T \propto N^{5+\delta}, \quad \delta \in [0, 1] \quad (6)$$

which is huge and clearly denies the analysis of even medium-sized graphs ($\mathcal{O}(10^4)$ nodes) without further improvements. Note that graphs of thousands of nodes may be manageable, like the word association graph shown above, which took 56 hours on a single CPU. One possibility is to choose the initial seed more intelligently, starting communities from promising seeds. [63] achieved good results in this aspect. An intelligent seed selection is also important if the number

of communities in a cover is larger than N , or if some communities have only overlapping nodes – in this case, it may happen that all growth processes miss a certain community.

Other important question is the applicability of an advanced eigenvalue solver. Arpack++ [65] and SLEPc [66] were tried. The experience was that – despite their good asymptotic performance in the large matrix limit – for the occurring several small subgraphs the overhead of these complicated machineries was so large that made the final running time much higher than those obtained with the QR-decomposition algorithm.

Doing optimization in a multi-parameter space is a nontrivial task, because different parameters can lie in different ranges. Therefore, an important direction for future research is to investigate the best combination of the parameters in the fitness function, based on the evaluation of empirical data.

Finally, filtering the relevant communities from the found ones is also a challenging task. The natural approach is to apply statistical significance, which should be applied even if filtering was not needed. However, deciding the threshold significance value is not necessarily trivial in all cases. Furthermore, the current form of statistical significance accounts only for the separation of the community, not for its internal cohesion. This manifests itself e.g. in the low score of the almost-full-clique subgraph in the Zachary karate club (the dark purple group on Fig. 3). As the main advantage of the fitness function of eq. 4 is the inclusion of cohesion, it would be important to develop a statistical significance taking it into account.

Conclusions. The community detection problem currently suffers from two fundamental deficiencies. First, there is no definition of community which is precise enough to allow constructing community finding methods. Second, thorough testing a proposed algorithm is problematic, not independently from the previous difficulty. I attempted to improve both issues.

In this paper, I proposed a formal list of required properties for locally dense subgraphs, taking a step towards an applicable definition of the term “community”. Two properties, external separation and internal cohesion (“shape”) were named. External separation has already been applied by some of the community detection methods, and also by benchmarks. Internal cohesion was not considered explicitly earlier. No current method was found which satisfactorily applies both criterions. I demonstrated on simple examples that both properties are necessary; discarding either of them leads to counterintuitive results. Beyond allowing to construct new methods, these two criterions can also be used as a basis for testing existing ones. They also allow the characterization of a community by two independent quantities, instead of a single scalar.

I proposed a new composite fitness function which takes the two criterions into account. For the quantification of the internal cohesion, the second eigenvalue of the Laplacian matrix is applied, which provides appropriate results on characteristic graphs like cliques or chains. I also proposed a heuristic, by redesigning the LFK heuristic [39], which can find overlapping locally dense subgraphs of all scales, producing much less output than multiresolution methods but with less restrictions than imposed by assuming a hierarchical structure. Runs on the Zachary network and LFR benchmarks showed that the method is able to provide the expected results. Overlapping communities can be detected especially efficiently, similarly to other LFK-based heuristics [63]. However, significant improvements are yet to be implemented; e.g. reducing the running

time, finding a more effective filtering procedure for the output, or fine-tuning the relative weight of the separation and the cohesion terms in the fitness function.

7 Acknowledgments

I wish to thank János Kertész for several useful suggestions. I also thank the Eötvös University for the access to its HPC cluster, and Andrea Lancichinetti, who proposed the Arpack++ package and was extremely helpful about his software. Thanks are due to the authors of the OSLOM [34], LFR benchmark [25], overlapping mutual information [39], Radatools [67] and linegraph-creator [53] softwares for making their code publicly available. I am grateful for the referees for several comments which significantly improved the manuscript. Financial support from EU's 7th Framework Program's FET-Open to ICTeCollective project no. 238597 is acknowledged.

Appendix

A Second eigenvalue of two weakly connected cliques

Assume two cliques of n nodes, edge weights are 1. The two cliques are attached by a single edge having weight ϵ . Then the eigenvalue equations for the Laplacian matrix are

$$\sum_{\substack{j < n \\ j \neq i}} x_j + x_n - (n - 1 + \lambda)x_i = 0 \quad \forall i < n \quad (7)$$

$$\sum_{\substack{k > n \\ k \neq i}} x_k + x_{n+1} - (n - 1 + \lambda)x_i = 0 \quad \forall i > n + 1 \quad (8)$$

$$\sum_{j < n} x_j + \epsilon \cdot x_{n+1} - (n - 1 + \epsilon + \lambda)x_i = 0 \quad i = n \quad (9)$$

$$\sum_{k > n+1} x_k + \epsilon \cdot x_n - (n - 1 + \epsilon + \lambda)x_i = 0 \quad i = n + 1 \quad (10)$$

Adding the last two equations gives

$$\sum_j x_j - x_n - x_{n+1} + \epsilon x_n + \epsilon x_{n+1} - (n - 1 + \epsilon + \lambda)x_n - (n - 1 + \epsilon + \lambda)x_{n+1} = 0 \quad (11)$$

The eigenvector corresponding to the first eigenvalue (which is zero) is the constant vector, therefore for all other eigenvectors the sum of components should be zero in order to be orthogonal to the first one. Consequently $\sum_j x_j = 0$. Applying this and a minimal algebra results

$$x_n(\lambda + n) + x_{n+1}(\lambda + n) = 0 \quad (12)$$

$$x_n = -x_{n+1} \quad \text{if } \lambda \neq -n \quad (13)$$

If $\lambda = -n$ then eqs. 7-8 reduce to $\sum_{j < n} x_j = 0$ and $\sum_{k > n+1} x_k = 0$. Now consider the eigenspace corresponding to $\lambda = -n$, and look for eigenvectors such that $x_n = x_{n+1} = c$, $\sum_{j < n} x_j = -c$, $\sum_{j > n+1} x_j = -c$. In this eigenspace the number of free parameters are $1 + 2 \cdot (n - 2)$, corresponding to c and $x_1 \dots x_{n-1}, x_{n+2} \dots x_{2n}$ with two constraints. Altogether the dimension of the eigenspace (the multiplicity of $\lambda = -n$) is $2n - 3$. Adding the $\lambda = 0$ case, we are left with at most two unknown eigenvalues.

For $\lambda \neq -n$, we look for the solutions in the form $(a, \dots, a, b, -b, -a, \dots, -a)^T$. Then the eigenvalue equations are

$$(n - 2)a + b - (n - 1 + \lambda)a = 0 \quad (14)$$

$$(n - 1)a - \epsilon \cdot b - (n - 1 + \epsilon + \lambda)b = 0 \quad (15)$$

After simplifications,

$$-(1 + \lambda)a + b = 0 \quad (16)$$

$$(n - 1)a - (n - 1 + 2\epsilon + \lambda)b = 0 \quad (17)$$

Expressing λ from these equations reads

$$\lambda = \frac{b}{a} - 1 \quad (18)$$

$$\lambda = -(n-1+2\epsilon) + (n-1)\frac{a}{b} \quad (19)$$

Writing $\lambda = \lambda$ results

$$-n+1-2\epsilon + (n-1)\frac{a}{b} = \frac{b}{a} - 1 \quad (20)$$

$$(-n+1-2\epsilon)\frac{b}{a} + (n-1) = \left(\frac{b}{a}\right)^2 - \frac{b}{a} \quad (21)$$

Introducing $x = b/a$ gives

$$-x^2 + (-n+2-2\epsilon)x + (n-1) = 0 \quad (22)$$

$$x_{1,2} = \frac{n-2+2\epsilon \pm \sqrt{(n-2+2\epsilon)^2 + 4(n-1)}}{-2} \quad (23)$$

The term under the radical symbol can be approximated using the first two terms of the Taylor series $\sqrt{1-x} \approx 1 - x/2$

$$\sqrt{\dots} = \sqrt{(n+2\epsilon)^2 \left(1 - \frac{8\epsilon}{(n+2\epsilon)^2}\right)} \approx \quad (24)$$

$$\approx (n+2\epsilon) \left(1 - \frac{4\epsilon}{(n+2\epsilon)^2}\right) \quad (25)$$

which gives

$$x_1 \approx 1 - \frac{2\epsilon}{n+2\epsilon} \quad (26)$$

$$x_2 \approx 1 - (n+2\epsilon) + \frac{2\epsilon}{n+2\epsilon} \quad (27)$$

which, using eq. 18, leads to

$$\lambda_1 \approx -\frac{2\epsilon}{n+2\epsilon} \quad (28)$$

$$\lambda_2 \approx -(n+2\epsilon) + \frac{2\epsilon}{n+2\epsilon} \quad (29)$$

meaning that the last two eigenvalues of the Laplacian are found.

B Review of current methods

Here, a one-by-one review of methods follows, from the point of view of the separation & cohesion criterions.

Separation-targeted methods

Method of Lancichinetti et al. (LFK) [39] – although being a multiresolution method, it is informative to take a look at it with the resolution parameter (see eq. 1) fixed at $\alpha = 1$. Then the fitness function of a community, which is to be optimized, is simply the sum of in-degrees divided by the sum of degrees of the community members. Thus, this method is a clear implementation of the separation criterion. Consequently, it is not sensitive to the internal distribution of edges (Fig. 1a and 1b get the same fitness). The cohesion criterion is absent, so one clique on Fig. 1a has lower fitness than the union of the two cliques.

Labelpropagation [40] – the communities are defined as sets of nodes such that every node should belong to the community to which the majority of their neighbors do. Labelpropagation does not qualify the communities, just finds partitions obeying the majority rule. Consequently Fig. 1a can be judged as a proper single community, and Fig. 1b can be split by collecting each second node to the same cluster.

Infomap [18] – Infomap aims to minimize the length of the description of a random walk, using clusters. The best description length corresponds to the best trade-off between small cluster sizes (understood in in-degrees) and few links between clusters. It is straightforward to calculate that for the configuration on Fig. 1a, Infomap will properly separate the two cliques unless the number of inter-community links is larger than $6.9 \cdot 10^7$. Although this resolution limit looks practically unimportant, shows that Infomap has some conceptual problems. If 3 edges are placed instead of 1 between the 2 cliques on Fig. 1a, Infomap will merge the two cliques if the number of inter-community edges in the rest of the network is larger than 149, which is more than 5 orders of magnitude smaller than the previous threshold. Two consequences should be drawn: Infomap is quite sensitive to the number of inter-community edges, and, as a consequence, it can produce counterintuitive communities in realistic graphs.

Clique Percolation Method (CPM) [21] – communities are defined as maximal sets of adjacent k -cliques, k being a parameter. Adjacency holds if $k - 1$ nodes are shared by two cliques. Although CPM enforces a very strong cohesion locally, it applies only to $\mathcal{O}(1)$ -sized subgraphs of communities. Consequently, there are no cohesion requirements on the scale of the whole community. E.g., the cliques of a cluster might form a chain and the method gives no information about the shape of the cluster. Considering Fig. 1a, it is trivial to modify it such that CPM merges the two large cliques into a single cluster, e.g. using 3-cliques. Furthermore, the absence of a single percolating series of neighboring cliques means that a subgraph will not appear as a single community, regardless of its other parameters (see e.g. Fig. 1b applying 4-cliques). Finally, CPM uses the same clique size for the whole network, regardless of local variations in edge density.

Method of Radicchi et al. [20] – there are two possible criterions for communities to choose from: either all community members or only the whole community should have more links inside than outside. Proper communities are found

by iteratively bisecting the network, until no bisection can be carried out without violating the criterion used. So, the effective definition is that a community is a subgraph obeying one of the criteria mentioned above *such that no bisection of it can result proper communities*. Fig. 1b with a minor tweak would be split even using the strong definition, assigning every second node to the same community. The tweak is to place the 2 outside links on the $k_{in} = 6$ nodes.

Method of Estrada and Hatano [41] – as it relies on the eigenvalues and eigenvectors of the whole graph, it is a global method. Therefore whether a set of nodes is judged to be a cluster or not depends also on the rest of the graph. Unfortunately, the behavior of the eigenvalues and eigenvectors of the adjacency matrix of a graph are not well understood. Consequently, empirical tests were conducted. If the method is run on only the 12 nodes of Fig. 1, configuration a) is cut into the two proper sets, but configuration b) is cut into several small (overlapping) clusters, such that all triangles form one. When the 12 nodes are attached to a 100-node ring, in which first and second neighbors on both sides of a node are attached to the node (degrees are 4), then for configuration a) the two clusters expand to the first neighboring nodes in the ring, and configuration b) has the same clusters as in the fully separated case. So, if the rest of the graph is not denser than the set of nodes under investigation, it seems that internal cohesion does matter, however external separation not. If the 100-node ring is two degrees denser (first 3 neighbors are attached, degrees are 6), the 12 nodes coalesce into 1 cluster both for configurations a) and b), incorporating a few nearby nodes from the large ring (8 for a) and 6 for b)). For even denser 100-node rings, the 12 nodes become part of a large cluster containing many nodes from the large ring. So, in conclusion, the global character of the method makes it indefinite concerning its behavior to the configurations on Fig. 1a and 1b.

Stochastic blockmodels and spin-based methods

Modularity optimization [17] – for each community, modularity counts the inside links and their expected values, based on the degrees of the nodes. Due to the well-known resolution limit problem [35, 36], the optimal modularity merge the two cliques on Fig. 1a for sufficiently large graphs.

Laplacian spectral algorithm by Donetti and Muñoz [42] – although the method produces candidate partitions using the spectrum of the Laplacian matrix, the partitions are evaluated using modularity. Consequently, it is equivalent to modularity optimization using a special heuristic, implying all the drawbacks of modularity.

Link partitioning method of Evans and Lambiotte [53] – partitioning is done on the so-called line graph, which nodes correspond to the edges of the original graph, and links are drawn between edges sharing a node in the original graph. Variants of the modularity function are proposed as goal function for the partition. Different variants use different weighting schemes of the edges including the addition of self-loops. As these goal functions are still based on counting intra-community edges and subtracting some expected value, the resolution limit problem should appear for large enough graphs.

Method of Ronhovde and Nussinov (RN) [43] – it proposes a Hamiltonian $\mathcal{H}(\{\sigma\}) = -1/2 \sum_{i \neq j} (a_{ij} A_{ij} - \gamma b_{ij} (1 - A_{ij})) \delta(\sigma_i, \sigma_j)$, \mathbf{A} being the adjacency matrix, a_{ij} and b_{ij} being edge weights. The configuration corresponding to the

minimal Hamiltonian is used as the solution.

The Hamiltonian optimizes simply for the edge densities inside clusters (distorted by the γ resolution parameter), which tends to be the largest for cliques. Consequently Fig. 1b worth to be split into 4 if $\gamma > 19/35$. Similarly, for $\gamma < 1/35$, the two cliques of 1a are merged. The fact that the proper value of γ may vary from cluster to cluster can render the global optimization process locally unsuccessful.

Method of Nepusz et al. [44] – the main goal of the work is to provide community detection framework using fuzzy (soft) memberships, in order to handle overlaps. The proposed realization of the framework, when restricted to conventional hard memberships (and unweighted networks), is equivalent to the previous method with $\gamma = 1$, and with a different heuristic.

Stochastic blockmodel of Hofman and Wiggins [45] – based on the assumption that the community structure can be fitted by a blockmodel in which intra- and inter-cluster nodes are connected with probabilities ϑ_c and ϑ_d respectively, [45] aims to minimize the Hamiltonian

$$H = - \sum_{i < j} (J_L A_{ij} - J_G) \delta_{\sigma_i, \sigma_j} - \sum_{\mu} n_{\mu} \ln \frac{n_{\mu}}{n} \quad (30)$$

where σ_i is the cluster of node i , n_{μ} is the size of cluster μ , $J_G = \ln(1 - \vartheta_d)/(1 - \vartheta_c)$, $J_L = \ln \vartheta_c / \vartheta_d + J_G$. The number and sizes of clusters, the cluster members, and the probabilities ϑ_c and ϑ_d are determined by minimization. In other words, a community structure should be found which maximizes the edge densities inside the communities (the $J_L A_{ij} - J_G$ term), with the restrictions that 1) each node belongs to exactly one community; 2) the expected intra- and inter-cluster edge densities are both constants. The method of Hastings [46] is a special case of this method, needing J_G and J_L as input, and discarding the last term in equation 30. The method of Ronhovde and Nussinov [43] is also a special case with the same restrictions, i.e. being equivalent to the Hastings method. One can see immediately that equation 30 defines a global method, which is realized by the global J_G and J_L coupling constants. Since the discovery of the resolution limit of modularity it is known that globality leads to counterintuitive local trade-offs. The situation is not different here, a simple calculation for Fig. 1a shows that the two cliques will be merged if $(1 - \vartheta_c)/(1 - \vartheta_d) > 2^{-12/35}(\vartheta_d/\vartheta_c)^{1/35}$, which can be approximated by $1 - \vartheta_c > 0.79(1 - \vartheta_d)$, assuming that $(\vartheta_d/\vartheta_c)^{(1/35)} \approx 1$. As ϑ_c corresponds to the intra-cluster edge probability, it is a reasonable criterion. Furthermore, it is similarly simple to show that for Fig. 1b, splitting into four is profitable if $(1 - \vartheta_d)/(1 - \vartheta_c) > 4^{12/35}(\vartheta_c/\vartheta_d)^{19/35}$.

Mixture model of Newman and Leicht [47] – based on some probabilistic modeling, [47] proposed the following log-likelihood to be maximized:

$$\mathcal{L} = \sum_{i,r} q_{ir} \left(\ln \pi_r + \sum_j A_{ij} \ln \Theta_{rj} \right) \quad (31)$$

where q_{ir} is the probability that node i belongs to cluster r , π_r is the fraction of nodes in cluster r , and Θ_{rj} is the probability that a randomly chosen link originating in cluster r points to node j . Equation 31 is reminiscent of the Hamiltonian of Hofman and Wiggins, although there are important differences. Nodes can have memberships in many clusters simultaneously (with the

constraint that the sum of memberships is 1 for any node). Inter-cluster edges are counted for, while missing edges are never. The coupling strength between neighboring nodes is fine-tuned for each node-cluster pair. Considering Fig. 1b and assuming hard node memberships (i.e. q_{ir} is 0 or 1 for all nodes), it is easy to show that splitting into 4 is favored over putting all nodes into one cluster. **Mixture model of Wang and Lai [48]** – Wang and Lai improved the mixture model of Newman and Leicht, arriving to the log-probability

$$\mathcal{L} = \sum_{i,r} q_{i,r} \left(\ln \pi_r + \sum_j A_{ij} \ln \rho_{rj} + \sum_j (1 - A_{ij}) \ln(1 - \rho_{rj}) \right) \quad (32)$$

where ρ_{rj} is the probability that a node in cluster r has a link to node j . Now \mathcal{L} counts also the missing edges. For a hard clustering ($q_{ir} = 0$ or 1) it is easy to calculate that Fig. 1b is preferred in 4 pieces over 1.

Likelihood modularity of Bickel and Chen [49] – the proposition is to maximize

$$Q_{LM} = \frac{1}{2} \sum_{c,d} n_{cd} \left(\frac{O_{cd}}{n_{cd}} \log \frac{O_{cd}}{n_{cd}} + \left(1 - \frac{O_{cd}}{n_{cd}} \right) \log \left(1 - \frac{O_{cd}}{n_{cd}} \right) \right) \quad (33)$$

where $n_{cd} = n_c n_d$ if $c \neq d$, $n_{cc} = n_c(n_c - 1)$, n_c is the size of cluster c , and $O_{cd} = \sum_{i \in c, j \in d} A_{ij}$. The expression is maximal if the clusters are cliques ($O_{cc}/n_{cc} = 1$) which are totally separated ($O_{cd}/n_{cd} = 0$). As Q_{LM} is symmetric with respect to O_{cd}/n_{cd} and $1 - O_{cd}/n_{cd}$, bipartite structures can also get high scores, but here the analysis is restricted to the cluster-based optimum. First, it should be noted that Q_{LM} penalizes clusters in which the edge density deviates significantly from its maximal value. Then, it is easy to calculate that it worth to cut Fig. 1b into 4 clusters.

Stochastic blockmodel of Karrer and Newman [50] – it is similar to the previous case. The main difference in the function to be maximized (compared to equation 33) is the absence of the second logarithmic term representing the missing links, and the application of sums of degrees instead of cluster sizes. Similarly, a simple calculation shows that Fig. 1b gets higher score when split into four.

Other single-scale methods

Infomod [51] – the aim is to compress the description of the graph, while retaining as much information as possible. The description length is given by

$$L = n \log_2 m + \frac{m(m+1)}{2} \log_2 l + \log_2 \prod_{i=1}^m \binom{n_i(n_i-1)/2}{l_{ii}} \prod_{i < j} \binom{n_i n_j}{l_{ij}} \quad (34)$$

where n is the number of nodes, l is the number of edges, m is the number of clusters. As can be seen, it is global method, where trade-offs for a global improvement may spoil local structures. And indeed, a straightforward calculation shows that for all but very small graphs Fig. 1a is preferred as a single community (e.g. $l \geq 128$ and $m \geq 7$).

Method of Chauhan et al. [52] – the idea is interesting, i. e. to maximize the sum of logarithms of the largest eigenvalues of the adjacency matrices of the

individual communities. However, the behavior of the largest eigenvalue of the adjacency matrix is poorly understood. As a counterexample, given a clique of size n , its largest eigenvalue is $n - 1$, while when it is cut into two, the product of the first eigenvalues of the two $n/2 - 1$ -sized cliques is $(n/2 - 1)^2$, which is larger than n if $n \geq 8$ – so it worth to cut a clique into pieces. This is the consequence of using a concave function (log) in the summation, so it can be easily fixed. However, for Fig. 1b, the largest eigenvalue is 5.2, while the largest eigenvalue of a 3-clique is 2. Summing the largest eigenvalues for the two cases (instead of summing their logarithms) results in $5.2 < 8$, so it worth to split Fig. 1b into 4.

Link partitioning method of Ahn et al. [54] – although the described method applies a hierarchical clustering, using an objective function (edge density of clusters) results in a single set of communities. The objective function averages the densities of all clusters, consequently it is a global quantity; its maximum does not guarantee that each cluster is optimal, just the average – nothing prevents the over- or underpartitioning of individual clusters at the global optimum.

Community landscape method of Kovács et al. (ModuLand) [55] – see also at the hierarchical methods. The interesting idea is to give a scalar value to the edges indicating how strongly an edge belongs to communities, then identifying the local maxima & their surroundings (“hills”) as the communities. The scalar value for the edges is obtained as the number of appearances of the edges in some auxiliary clusters. From each node (or edge), an auxiliary cluster is grown until its fitness value cannot be increased. Fitness is chosen as simply the average in-degree of the nodes in the growing cluster. After all auxiliary communities are determined this way, each edge is assigned a value equaling the number of times it occurred in the found communities. Edges with the locally highest score are defined as community cores. Membership values are assigned to remaining edges, based on how strongly are they related to the nearby cores. The method, actually a framework for several possible methods, depends heavily on the applied fitness function of the auxiliary clusters. Here the NodeLand auxiliary clustering will be investigated. It is quite easy to engineer graphs in the spirit of Fig. 1 which are misclustered. E.g instead of Fig. 1a, took two 7-cliques, delete 1 link from each, and connect one node with 3 nodes from the other clique, as on Fig 7a. The fitness of one almost-clique is $40/7$, is just below the contribution of the node in the other clique ($6/1$), so the 3 links between the cliques will be included in the community. To be precise, starting a cluster from each node, one almost-clique + the connector node will appear as a cluster $7 + 1/3$ times, and the other almost-clique $6 + 2/3$ times. Fractions correspond to different possibilities when starting from the connector node. In practice, this means that with probability $2/3$, all links will have uniform scalar values (perfectly flat landscape, i.e. a single hilltop), and with probability $1/3$, a step-like landscape (still identified as a single cluster by the method). Symmetrization to $6 + 1/3 + 2/3$ and $6 + 2/3 + 1/3$ is straightforward, by creating a second bridge node also with 3 links. Similarly, Fig. 1b can be substituted by Fig. 7b. It consists of two 5-cliques with connections such that each node has 2 links to the other clique. ModuLand-NodeLand will tend to separate the two cliques, although their union is much more well-separated from the rest of the graph.

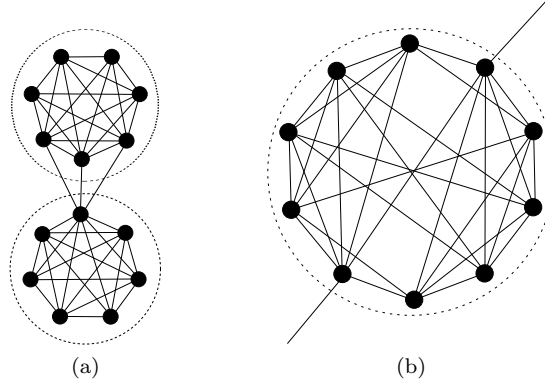


Figure 7: Subgraphs on which ModuLand-NodeLand gives counterintuitive results. Dashed lines show the desired communities.

Hierarchical methods

Here, some hierarchical methods will be investigated. The question is whether the lowest level can be reliably used as an optimal partition (or cover). As a benchmark graph, Fig. 8 will be utilized. The desired output is a single community of 12 nodes, due to their extreme separation from the rest of the graph.

Method of Ruan and Zhang [68] – the proposition is to iteratively run

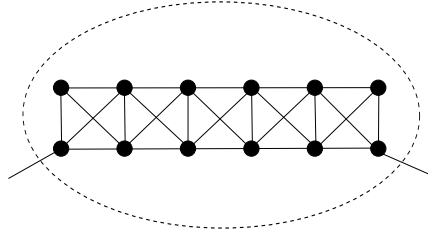


Figure 8: Test case for the lowest level of the hierarchical methods.

modularity optimization in the found clusters, until the best modularity inside a cluster is not larger significantly than those of a corresponding random graph. Numerical calculations show that at the lowest level, Fig. 8 is divided into parts¹².

Method of Sales-Pardo et al. [69] – it uses the co-occurrence of nodes in different local optima of modularity to construct a new similarity matrix, which is fitted by a block diagonal form. Communities are defined by the blocks. The method is iteratively re-applied to each community until structure deviating from a corresponding random graph is found. Again, running the method on Fig. 8 results in overpartitioning (z-score of the split Fig. 8 is 3.9, the threshold

¹²z-score is 5.3, $Q_{\max} = 0.36$. Z-score is defined as the difference of the modularity of the actual graph and the modularity of a 0-model graph, divided by the variance of the modularity of the 0-model graph, $z\text{-score} = (Q - Q_{0\text{-model}})/\sigma_{0\text{-model}}$. Criterion of [68] is $z\text{-score} \geq 2$, $Q_{\max} \geq 0.3$. Modularities were optimized using the Radatools software [67].

used by [69] is 2.3).

Hierarchical Infomap [70] – this is an extension of the Infomap method [18]. It is easy to calculate that, similarly to the previous cases, splitting Fig. 8 on a lower hierarchical level improves the partition.

ModuLand [55] – ModuLand can also produce hierarchical structures, by iteratively re-running the clustering procedure on the network of clusters (links between clusters are defined by node overlaps). Accordingly, the lowest level clusters are the ones obtained by a simple ModuLand run, which is susceptible to mispartitioning, as described some paragraphs above.

OSLOM [34] – the method applies statistical significance as fitness. Although its output depends to a certain degree on the whole graph, running it on Fig. 8 (as the whole graph) results in a bisection. As the method tries to find the so-called minimal significant clusters, by trying to split already found significant subgraphs while the rest of the graph is neglected, it will divide Fig. 8 independently of the rest of the graph.

References

- [1] R. Albert, A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47-97 (2002).
- [2] M. E. J. Newman, *SIAM Review* **45**, 167–256 (2003).
- [3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. Hwang, *Physics Reports* **424**, 175-308 (2006).
- [4] M. Newman, A.-L. Barabási, D. Watts: *The Structure and Dynamics of Networks* (Princeton University Press, 2006).
- [5] G. Caldarelli, A. Vespignani (eds.): *Large Scale Structure and Dynamics of Complex Networks*, (World Scientific, 2007).
- [6] S. N. Dorogovtsev, J. F. F. Mendes: *Evolution of Networks: From Biological Nets to the Internet and WWW*, (New York: Oxford University Press, 2003).
- [7] D. J. Watts, S. H. Strogatz, *Nature* **393**, 440-442 (1998).
- [8] A.-L. Barabási, R. Albert, *Science* **286**, 509–512 (1999).
- [9] R. Albert, H. Jeong, A.-L. Barabási, *Nature* **401**, 130-131 (1999).
- [10] R. Cohen, S. Havlin, *Phys. Rev. Lett* **90**, 058701 (2003).
- [11] R. Pastor-Satorras, A. Vespignani, *Phys. Rev. E* **63**, 066117 (2001).
- [12] J. D. Noh, H. Rieger, *Phys. Rev. Lett.* **92**, 118701 (2004).
- [13] M. Barahona, L.M. Pecora, *Phys. Rev. Lett.* **89**, 054101 (2002).
- [14] S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes, *Rev. Mod. Phys.* **80**, 1275-1335 (2008).
- [15] C Hauert, G Szabó, *Am. J. Phys.* **73**, 405-414 (2005).

- [16] S. Fortunato, *Phys. Rep.* **486**, 75-174 (2010).
- [17] M. E. J. Newman, M. Girvan, *Phys. Rev. E* **69** 026113 (2004).
- [18] M. Rosvall, C. T. Bergstrom, *PNAS* **105** 1118-1123 (2008).
- [19] J.-C. Delvenne, S. N. Yaliraki, M. Barahona, *PNAS* **107**, 12755-12760 (2010).
- [20] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, *PNAS* **101** 2658-2663 (2004).
- [21] G. Palla, I. Derényi, I. Farkas, T. Vicsek, *Nature* **435** 814-818 (2005).
- [22] J. Reichardt, S. Bornholdt, *Phys. Rev. E* **74**, 016110 (2006).
- [23] A. Arenas, A. Fernández, S. Gómez, *New J. Phys.* **10**, 053039 (2008).
- [24] J. Kumpula, J. Saramäki, K. Kaski, J. Kertész, *Fluct. Noise Lett.* **7**, L209 (2007).
- [25] A. Lancichinetti, S. Fortunato, F. Radicchi, *Phys. Rev. E* **78**, 046110 (2008).
- [26] A. Lancichinetti, S. Fortunato, *Phys. Rev. E* **80**, 016118 (2009).
- [27] R. Guimerà, M. Sales-Pardo, L. A. N. Amaral, *Phys. Rev. E* **70**, 025101 (2004).
- [28] C. P. Massen, J. P. K. Doye, arXiv:cond-mat/0610077v1 (2006).
- [29] A. Lancichinetti, F. Radicchi, J.J. Ramasco, *Phys. Rev. E* **81**, 046110 (2010).
- [30] B. Karrer, E. Levina, M. E. J. Newman, *Phys. Rev. E* **77**, 046119 (2008).
- [31] Y. Hu, Y. Ding, Y. Fan, Z. Di, arXiv:1002.2007v1 (2010).
- [32] M. Rosvall, C. T. Bergstrom, *PLoS ONE* **5**, e8694 (2010).
- [33] D. Gfeller, J.-C. Chappelier, P. De Los Rios, *Phys. Rev. E* **72**, 056135 (2005).
- [34] A. Lancichinetti, F. Radicchi, J. J. Ramasco, S. Fortunato, *PLoS ONE* **6**, e18961 (2011).
- [35] S. Fortunato, M. Barthélemy, *PNAS* **104** 36-41 (2007).
- [36] J.M. Kumpula, J. Saramäki, K. Kaski, J. Kertész, *Eur. Phys. J. B* **56**, 41-45 (2007).
- [37] A. Lancichinetti, M. Kivela, J. Saramäki, S. Fortunato, *PLoS ONE* **5**, e11976 (2010).
- [38] G. Tibély, M. Karsai, L. Kovanen, K. Kaski, J. Kertész, J. Saramäki, *Phys. Rev. E* **83**, 056125 (2011).
- [39] A. Lancichinetti, S. Fortunato, J. Kertész, *New J. Phys.* **11** 033015 (2009).

- [40] U. N. Raghavan, R. Albert, S. Kumara, *Phys. Rev. E* **76** 036106 (2007).
- [41] E. Estrada, N. Hatano, *Phys. Rev. E* **77**, 036111 (2008).
- [42] L. Donetti, M. A. Muñoz, *J. Stat. Mech.* P10012 (2004).
- [43] P. Ronhovde and Z. Nussinov, *Phys. Rev. E* **80**, 016109 (2009).
- [44] T. Nepusz, A. Petróczi, L. Négyessy, F. Bazsó, *Phys. Rev. E* **77** 016107 (2008).
- [45] J. M Hofman, C. H. Wiggins, *Phys. Rev. Lett.* **100**, 258701 (2008).
- [46] M. B. Hastings, *Phys. Rev. E* **74** 035102 (2006).
- [47] M. E. J. Newman, E. A. Leicht, *PNAS* **104** 9564-9569 (2007).
- [48] J. Wang, C-H Lai, *New J. Phys.* **10**, 123023 (2008).
- [49] P. J. Bickel, A. Chen, *PNAS* **106**, 21068-21073 (2009).
- [50] B. Karrer, M. E. J. Newman, *Phys. Rev. E* **83**, 016107 (2011).
- [51] M. Rosvall, C. T. Bergstrom, *PNAS* **104** 7327-7331 (2007).
- [52] S. Chauhan, M. Girvan, E. Ott, arXiv:0911.2735v1 (2009).
- [53] T. Evans, R. Lambiotte, *Phys. Rev. E* **80**, 016105 (2009).
- [54] Y.-Y. Ahn, J. P. Bagrow, S. Lehmann, *Nature* **466**, 761-764 (2010).
- [55] I. A. Kovács, R. Palotai, M. S. Szalay, P. Csermely, *PLoS ONE* **5**, 12528 (2010).
- [56] A. Clauset, C. Moore, M. E. J. Newman, *Nature* **453**, 98-101 (2008).
- [57] S. van Dongen, Ph.D. thesis, Dutch National Research Institute for Mathematics and Computer Science, University of Utrecht, Netherlands (2000).
- [58] B. Mohar, S. Poljak, *Combinatorial and Graph-Theoretical Problems in Linear Algebra* **50**, 107-151 (1993).
- [59] M. Fiedler, *Czechoslovak Math. J.* **23**, 298-305 (1973).
- [60] A. Clauset, *Phys. Rev. E* **72**, 026132 (2005).
- [61] W. W. Zachary, *J. Anthropol. Res.* **33**, 452-473 (1977).
- [62] A. Lancichinetti, S. Fortunato, *Phys. Rev. E* **80**, 056117 (2009).
- [63] C. Lee, F. Reid, A. McDaid, N. Hurley, conference paper, Workshop - ACM KDD-SNA (preprint: arXiv:1002.1827v1) (2010).
- [64] D. L. Nelson, C. L. McEvoy, T. A. Schreiber, "The university of south florida word association, rhyme, and word fragment norms" (1998) <http://www.usf.edu/FreeAssociation/>.
- [65] <http://www.ime.unicamp.br/~chico/arpack++/>

- [66] V. Hernandez, J. E. Roman, V. Vidal, *ACM Transactions on Mathematical Software* **31**, 351-362 (2005).
- [67] <http://deim.urv.cat/~sgomez/radatools.php>
- [68] J. Ruan, W. Zhang, *Phys. Rev. E* **77**, 016104 (2008).
- [69] M. Sales-Pardo, R. Guimerà, A. A. Moreira, L. A. N. Amaral, *PNAS* **104**, 15224-15229 (2007).
- [70] M. Rosvall, C. T. Bergstrom, *PLoS ONE* **6**, e18209 (2011).